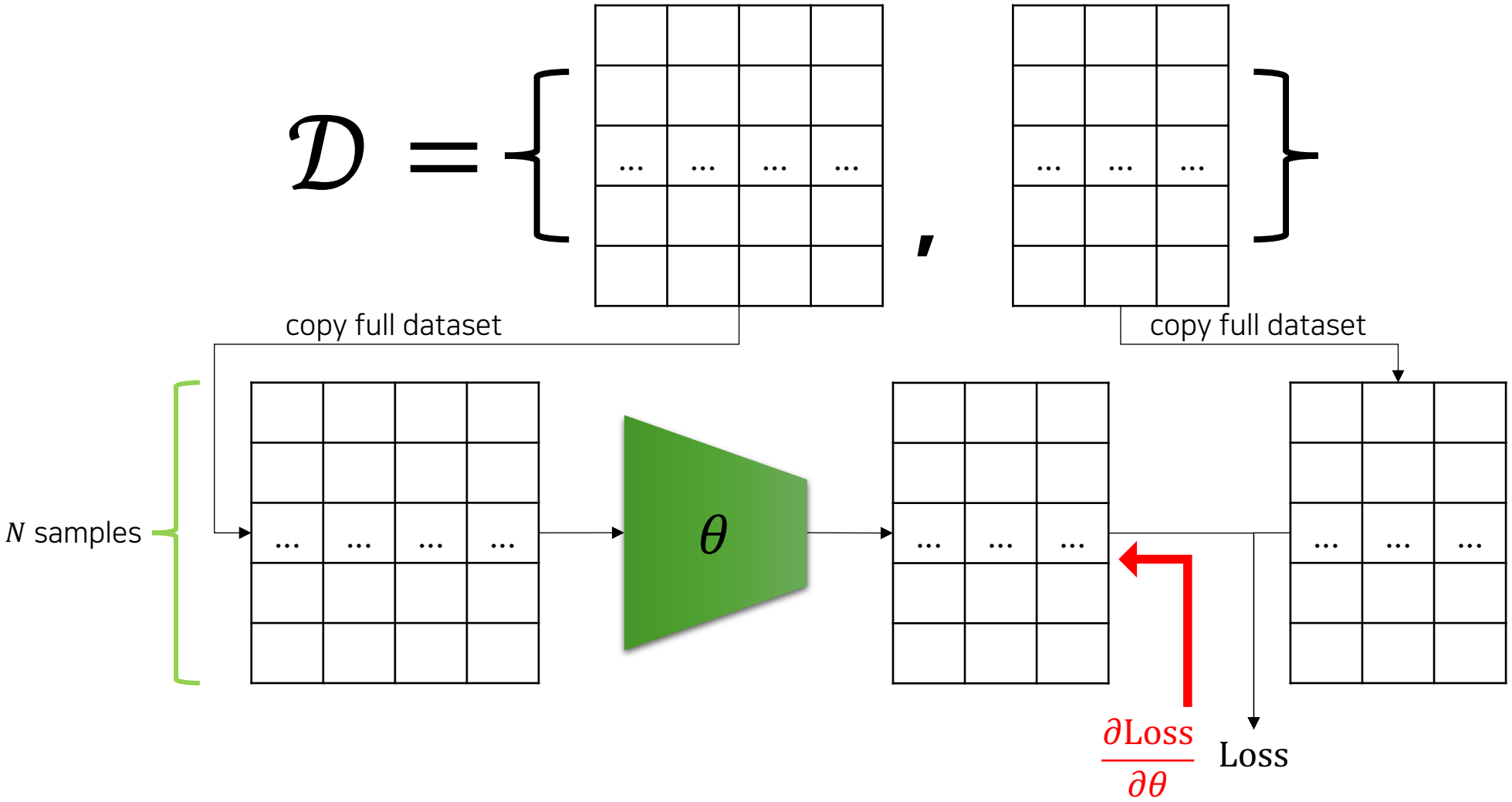
Stochastic Gradient Descent (SGD)

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

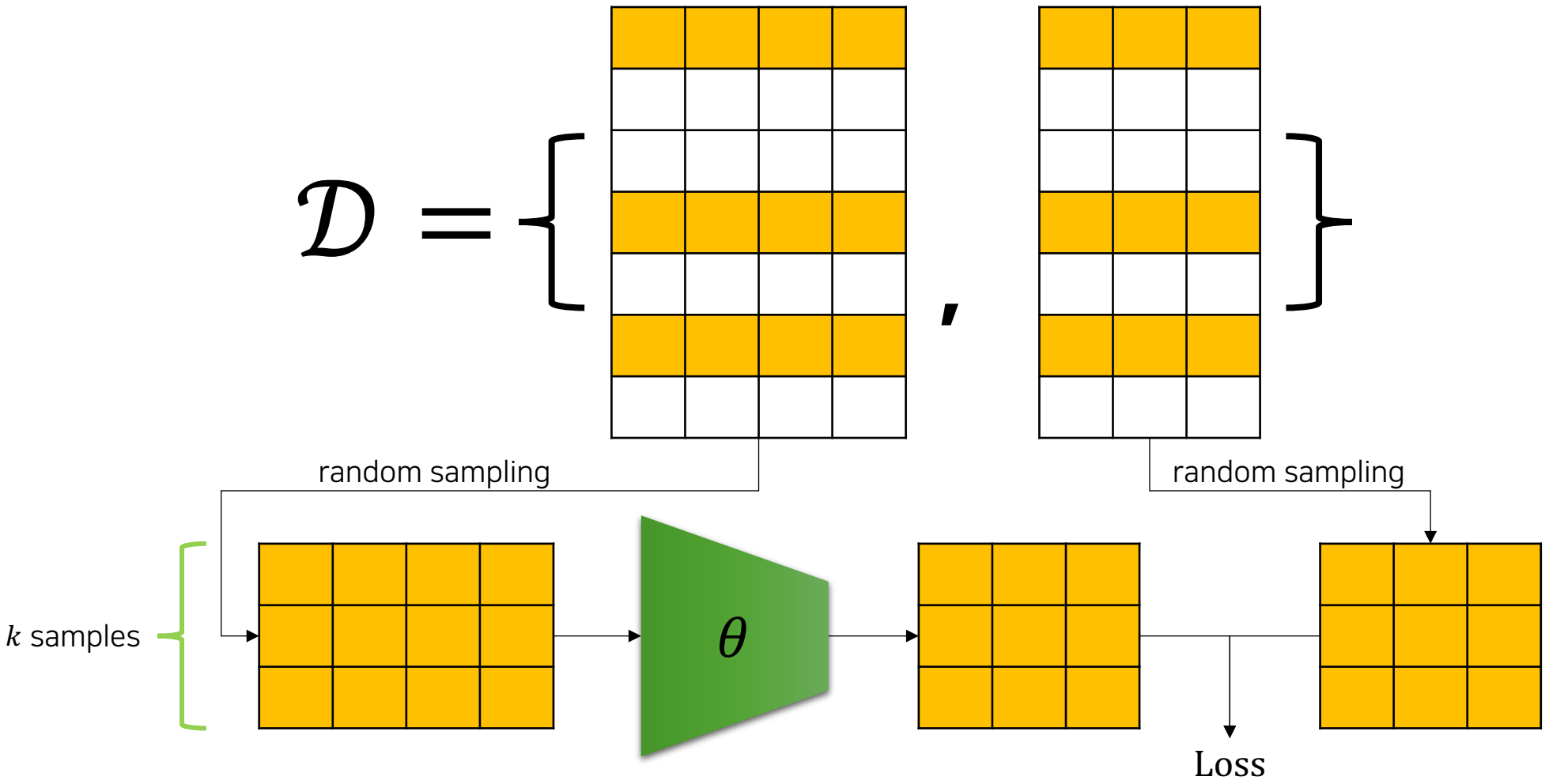
Currently, what we do

- 1 parameter update by GD from full sample's loss. → too expensive



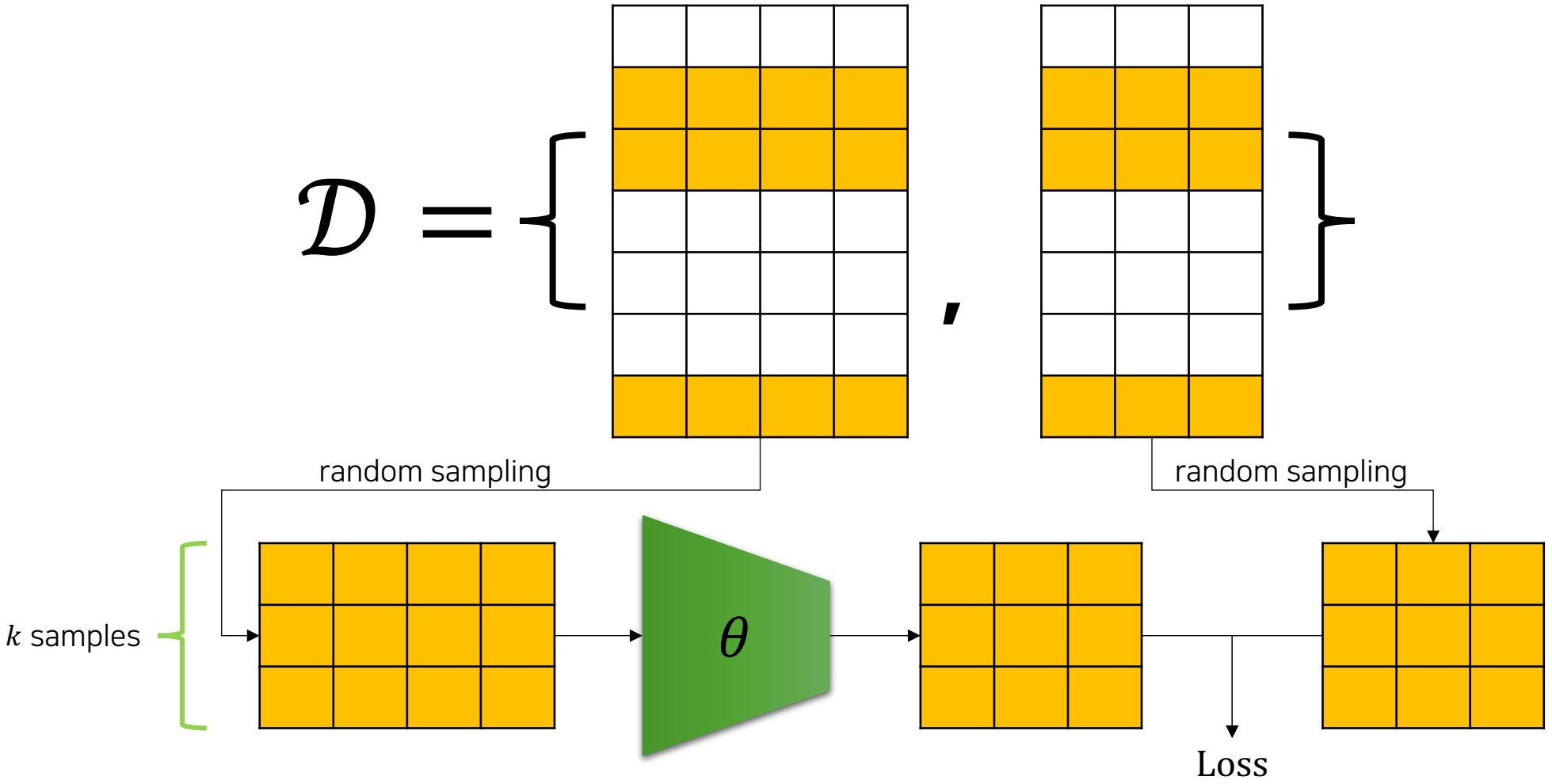
Stochastic Gradient Descent (SGD)

- 1st update from random k sample loss.



Stochastic Gradient Descent (SGD)

- 2nd update from another random k sample loss.



Epoch & Iteration

- 1 Epoch
  - 모든 데이터셋의 샘플들이 forward & backward 되는 시점
  - Epoch의 시작에 데이터셋을 random shuffling 해준 후, 미니배치로 나눈다.

- 1 Iteration
  - 한 개의 미니배치 샘플들이 forward & backward 되는 시점

- 따라서 Epoch와 Iteration의 이중의 for loop이 만들어지게 됨
  - 파라미터 전체 업데이트 횟수: $\#\text{epochs} \times \#\text{iterations}$

SGD Summary

- 전체 샘플의 loss에 대한 gradient descent가 아닌,
  일부 샘플($k = \text{batch\_size}$)의 loss에 대한 gradient descent.

- 1 epoch(전체 데이터셋을 활용한 학습)당 파라미터 update 횟수 증가
  - 대신 1 epoch의 소요 시간 증가

- batch_size가 작아질수록 gradient가 실제 gradient와 달라질 것
  - *어쩌면* 이로 인해 local minima를 탈출 할 수도 있음

- 요즘 학계의 추세는 큰 batch_size를 가져가려함
  - GPU를 사용하면 병렬 연산으로 인해, 큰 배치사이즈로 인한 비용이 줄어들기 때문
    - e.g. 2048, 4096 등..
  - 매우 큰 배치사이즈의 경우에는 오히려 성능을 악화시킬 수도 있음